A Strong Law of Large Numbers for Strongly Mixing Processes

L. Kontorovich and A. Brockwell

July 29, 2008

Abstract

We prove a strong law of large numbers for a class of strongly mixing processes. Our result rests on recent advances in understanding of concentration of measure. It is simple to apply and gives finite-sample (as opposed to asymptotic) bounds, with readily computable rate constants. In particular, this makes it suitable for analysis of inhomogeneous Markov processes. We demonstrate how it can be applied to establish an almost-sure convergence result for a class of models that includes as a special case a class of adaptive Markov chain Monte Carlo algorithms.

1 Introduction

The strong laws of large numbers (SLLNs) play a fundamental role in statistics. They assert the convergence of empirical averages to true expectations, and, under appropriate assumptions, ensure that inferences about persistent world phenomena become increasingly more valid as data accumulates. The various laws of large numbers date back at least to the publication in 1713 of Jakob Bernoulli's *Ars Conjectandi* (Bernoulli, 1713), which stated an early form of the weak law of large numbers. Subsequent development of the concept was carried out by many others. Among the first of these was Siméon-Denis Poisson, who generalized Bernoulli's result and gave it its modern name, "la loi de grands nombres". Other notable mathematicians who made further contributions include Chebyshev, Markov, Borel, Cantelli, and Kolmogorov. Over the years, the various contributions gave rise to two common forms, the weak and strong laws of large numbers, establishing conditions, respectively, for weak and strong convergence of empirical averages.

The SLLN in its earlier forms applies to sequences of independent random variables. However, for dependent sequences the theory is not as well developed. Indeed, to quote Ninness (2000), "for non-iid sequences the required SLLN results do not seem to be readily available in the literature" and thus researchers continue to work to develop generalized strong laws of large numbers. Since around 1960, a number of different generalizations have been obtained. Our goal is not to give a comprehensive survey of these results; the interested reader may find a useful list of relevant papers at `http://www.stats.org.uk/law-of-large-numbers/`. The basic current state of affairs may be summarized as follows. From Birkhoff's ergodic theorem we get a law of large numbers for ergodic processes; this has been strengthened by Breiman (1960) to cover the case where the stationary distribution is singular with respect to the Lebesgue measure. Assumptions of ergodicity

are typically too weak to provide a convergence rate – this requires a stronger mixing condition. A classical (and perhaps first of its kind) example of the latter is the paper by Hanson et al. (1963), which proves a strong law of large numbers under a mixing condition known in a modern form as ψ-mixing (see Bradley (2005)). This mixing condition guarantees exponentially rapid convergence, but the proof does not directly yield rate constants.

We approach the problem of developing such a SLLN with easily computable rate constants. This is useful for example, in the context of simulation-based algorithms, if one is interested in determining how many iterations are required to achieve a specified accuracy. To this end, we assume a stronger (though still quite realistic) mixing condition from which we obtain strong laws of large numbers, along with finite-sample bounds, for sequences of random variables with arbitrary dependence. In a sense, the application of concentration of measure theory to establish a SLLN is straightforward, although an important technical contribution of this paper is to extend the deviation bounds proved in Kontorovich and Ramanan (2008); Kontorovich (2006b,c) for discrete measure spaces to the continuous case, as well as to bound certain mixing coefficients for adaptive Markov chains and to prove a strong law of large numbers for these (Theorem 5.7). We will state our results in a fairly general setting. We emphasize that the basic concentration result presented and employed here (Theorem 3.6) is by no means the only one available. The recent result of Chazottes et al. (2007) yields a very similar inequality; earlier results along these lines appeared in Marton (1998). The interested reader may refer to authoritative and comprehensive surveys of concentration techniques, such as Ledoux (2001), Schechtman (2003), and Lugosi (2003) or to Chatterjee (2005) and Kontorovich (2007) for more recent developments.

We mention in passing that the strong laws we consider here are not *uniform laws*, the latter asserting almost sure convergence uniformly over some permissible (i.e., Glivenko-Cantelli[1]) class of sets. There has been some work on such uniform laws for non-i.i.d. processes. In particular, Nobel and Dembo (1993) show that if $\mathcal{F}$ is a Glivenko-Cantelli class for an i.i.d. process, then the corresponding uniform strong law also holds for β-mixing (absolutely regular) process. In the other direction, Nobel (1995) gives a counterexample where the uniform strong law fails for a stationary ergodic process. These two papers give ample background and provide illuminating discussions; they are an excellent starting point for anyone wishing to delve deeper into the topic.

This paper is organized as follows. In Section 2 we set down the notations and definitions used throughout the paper. This is followed by Section 3, where we describe the martingale method as our main workhorse for proving concentration of measure results. Our main law of large numbers is stated in Section 4. This strong law of large numbers is applied to adaptive Markov chains in Section 5. Finally, some technical lemmas are deferred to the Appendix.

2 Notation and Definitions

Let $(\Omega^n, \mathcal{F}^n, \mathbf{P})$ be a probability space, where $\mathcal{F}^n$ is the usual Borel σ-algebra generated by the finite dimensional cylinders. On this space define the random process $(X_i)_{1\leq i\leq n}$, $X_i \in \Omega$. Throughout

[1] Following Pollard (1984), the term *permissible* is used to avoid measure-theoretic pathologies associated with taking suprema over uncountable collections of sets.

this paper, we assume $\mathbf{P} \ll \mu^n$, for some positive Borel product measure $\mu^n = \mu \otimes \mu \otimes \ldots \otimes \mu$ on $(\Omega^n, \mathcal{F}^n)$.

We use the indicator variable $\mathbb{1}_{\{A\}}(\omega)$, equal to 1 if $\omega \in A$ and 0 otherwise, for $\omega \in \Omega^n$. For brevity, we will occasionally omit the argument ω. The ramp function is defined by $(z)_+ = z\mathbb{1}_{\{z>0\}}$. For arbitrary Ω, we define the *Hamming* metric on Ω^n:

$$d_{\mathrm{Ham}}(x,y) \quad = \quad \sum_{i=1}^{n} \mathbb{1}_{\{x_i \neq y_i\}}, \qquad x, y \in \Omega^n. \tag{1}$$

We need to introduce the notion of η-mixing, defined in Kontorovich (2006c)[2]. We note that this type of mixing is by no means new; it can be traced (at least implicitly) to Marton (1998) and is quite explicit in Samson (2000) and Chazottes et al. (2007). For $1 \le i < j \le n$ and $x \in \Omega^i$, let

$$\mathcal{L}(X_{j:n} \,|\, X_{1:i} = x)$$

be the law (distribution) of $X_{j:n}$ conditioned on $X_{1:i} = x$. For $y \in \Omega^{i-1}$ and $w, w' \in \Omega$, define

$$\eta_{ij}(y, w, w') \quad = \quad \left\| \mathcal{L}(X_{j:n} \,|\, X_{1:i} = (y,w)) - \mathcal{L}(X_{j:n} \,|\, X_{1:i} = (y,w')) \right\|_{\mathrm{TV}}, \tag{2}$$

where for signed measures ν, $\|\nu\|_{\mathrm{TV}} = [\sup_A \nu(A) - \inf_A \nu(A)]/2$ is the total variation norm and

$$\bar{\eta}_{ij} \quad = \quad \operatorname*{ess\,sup}_{y \in \Omega^{i-1}, w, w' \in \Omega} \eta_{ij}(y, w, w'),$$

the essential supremum being taken with respect to μ.

Remark 2.1. The definition in (2) is flawed as stated, as it does not adequately handle the case where sets of measure zero are conditioned on. A more precise definition is given in Kontorovich (2008). Taking $\mathcal{P}_+^n(\Omega)$ to be the set of all strictly positive probability measures μ on Ω^n (i.e., $\mu(x) > 0$ for all $x \in \Omega^n$), it is not hard to show that the functional $\bar{\eta}_{ij} : \mathcal{P}_+^n(\Omega) \to \mathbb{R}$ is continuous on $\mathcal{P}_+^n(\Omega)$ with respect to $\|\cdot\|_{\mathrm{TV}}$. However, this continuity can break down on the boundary of $\mathcal{P}_+^n(\Omega)$. Thus, we may arbitrarily define $\bar{\eta}_{ij}$ on any probability measure μ by

$$\bar{\eta}_{ij}(\mu) \quad = \quad \inf_{\{\mu_k\}} \lim_{k \to \infty} \bar{\eta}_{ij}(\mu_k) \tag{3}$$

where the infimum is taken over all sequences $\left\{\mu_k : \mu_k \in \mathcal{P}_+^n(\Omega), \|\mu_k - \mu\|_{\mathrm{TV}} \to 0\right\}$.

See Section 5.4 of Kontorovich (2007) for a discussion of the continuity of $\bar{\eta}$ and conditioning on sets of measure zero, as well as motivation for the definition in (3).

Let Δ_n be the upper-triangular $n \times n$ matrix defined by $(\Delta_n)_{ii} = 1$ and

$$(\Delta_n)_{ij} = \bar{\eta}_{ij}. \tag{4}$$

for $1 \le i < j \le n$. Recall that the ℓ_∞ operator norm is given by

$$\|\Delta_n\|_\infty \quad = \quad \max_{1 \le i < n} (1 + \bar{\eta}_{i,i+1} + \ldots + \bar{\eta}_{i,n}). \tag{5}$$

[2]This notion of mixing is distinct from, and not to be confused with η-weak dependence of Doukhan and Louhichi (1999).

We will occasionally want to make the dependence of Δ_n on the probability measure P explicit; to do this we will write $\Delta_n(P)$. Let us collect some simple observations about $\|\Delta_n(\cdot)\|_\infty$:

Lemma 2.2. *Let $P \ll \mu^n$ be a probability measure on $(\Omega^n, \mathcal{F})$. Then*

(a) $1 \le \|\Delta_n(P)\|_\infty \le n$

(b) $\|\Delta_n(P)\|_\infty = 1$ *iff P is (equivalent a.e. $[\mu^n]$ to) a product measure*

(c) if $Q \ll \mu^m$ is a probability measure on Ω^m then

$$\|\Delta_{m+n}(P \otimes Q)\|_\infty \quad \le \quad \max\left\{\|\Delta_n(P)\|_\infty, \|\Delta_m(Q)\|_\infty\right\}.$$

These properties are established in Kontorovich (2007), which also gives a discussion of the relationship between η-mixing and ϕ- and other kinds of mixing.

3 Concentration via Martingale Differences

Recall our probability space $(\Omega^n, \mathcal{F}^n, \mathbf{P})$ and let $\mathcal{F}_i$ be the σ-algebra generated by $(X_1 \dots X_i)$, which induces the filtration

$$\{\emptyset, \Omega^n\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_n = \mathcal{F}^n. \tag{6}$$

For $i = 1, \ldots, n$ and $f \in L_1(\Omega^n, \mathbf{P})$, define the martingale difference

$$V_i \quad = \quad \mathbf{E}[f \,|\, \mathcal{F}_i] - \mathbf{E}[f \,|\, \mathcal{F}_{i-1}]. \tag{7}$$

It is a classical result[3], going back to Azuma (1967), that

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \le \quad 2\exp\left(\frac{-t^2}{2\sum_{i=1}^n \|V_i\|_\infty^2}\right). \tag{8}$$

Thus, if we are able to uniformly bound the martingale difference,

$$\max_{1 \le i \le n} \|V_i\|_\infty \quad \le \quad H_n,$$

we obtain the concentration inequality

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \le \quad 2\exp\left(-\frac{t^2}{2nH_n^2}\right). \tag{9}$$

Recall our assumption that $d\mathbf{P}(x) = p(x)d\mu^n(x)$ for some positive Borel product measure $\mu^n = \mu \otimes \mu \otimes \ldots \otimes \mu$ on $(\Omega^n, \mathcal{F})$. Similarly, the conditional probability satisfies $\mathbf{P}(\cdot \,|\, \mathcal{F}_i) \ll \mu^{n-i}$, with density

[3] See Ledoux (2001) for a modern presentation and a short proof of (8).

$p(\cdot \,|\, X_{1:i} = x_{1:i})$. Here and below $p(x_{j:n} \,|\, x_{1:i})$ will occasionally be used in place of $p(x_{j:n} \,|\, X_{1:i} = x_{1:i})$; no ambiguity should arise.

For $f \in L_1(\Omega^n, \mathbf{P})$, $1 \le i \le n$ and $y_{1:i} \in \Omega^i$, define

$$V_i(f; y_{1:i}) \quad = \quad \mathbf{E}[f(X) \,|\, X_{1:i} = y_{1:i}] - \mathbf{E}[f(X) \,|\, X_{1:i-1} = y_{1:i-1}]; \tag{10}$$

this is just the martingale difference.

A slightly more tractable quantity turns out to be

$$\hat{V}_i(f; y_{1:i-1}, w_i, w_i') \quad = \quad \mathbf{E}[f(X) \,|\, X_{1:i} = (y_{1:i-1}, w_i)] - \mathbf{E}[f(X) \,|\, X_{1:i} = (y_{1:i-1}, w_i')], \tag{11}$$

where $w_i, w_i' \in \Omega$. These two quantities have a simple relationship, which may be stated symbolically as

$$\|V_i(f; \cdot)\|_{L_\infty(\mathbf{P})} \quad \le \quad \|\hat{V}_i(f; \cdot)\|_{L_\infty(\mathbf{P})}; \tag{12}$$

this is proved in (Kontorovich, 2006c, Lemma 4.1).

The next step is to notice that $\hat{V}_i(\cdot; y_{1:i-1}, w_i, w_i')$, as a functional on $L_1(\Omega^n, \mathbf{P})$, is linear; in fact, it is given by

$$\hat{V}_i(f; y_{1:i-1}, w_i, w_i') \;=\; \int_{\Omega^n} f(x)\hat{g}(x) d\mu^n(x) \;=\; \langle f, \hat{g} \rangle \,, \tag{13}$$

where

$$\hat{g}(x) \quad = \quad \mathbb{1}_{\{x_{1:i} = (y_{1:i-1}, w_i)\}} p(x_{i+1:n} \,|\, (y_{1:i-1}, w_i)) - \mathbb{1}_{\{x_{1:i} = (y_{1:i-1}, w_i')\}} p(x_{i+1:n} \,|\, (y_{1:i-1}, w_i')). \tag{14}$$

The plan is to bound $\langle f, \hat{g} \rangle$ using continuity properties of f and mixing properties of X, which will immediately lead to a result of type (9) via (12).

3.1 Φ and Ψ Norms

To state our results in sufficient generality, we shall borrow several definitions from Kontorovich (2007). Let $(\mathcal{X}, \rho)$ be a metric space and recall the definition of the *Lipschitz constant* of an $f : \mathcal{X} \to \mathbb{R}$:

$$\|f\|_{\mathrm{Lip}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{\rho(x, y)}, \qquad x, y \in \mathcal{X}. \tag{15}$$

Recall also the definition of the *diameter*:

$$\mathrm{diam}_\rho(\mathcal{X}) \quad = \quad \sup_{x, y \in \mathcal{X}} \rho(x, y).$$

Let μ be a positive Borel measure on a measurable space $(\Omega, \mathcal{F})$ and let $F_n = L_1(\Omega^n, \mathcal{F}^n, \mu^n)$ be equipped with the inner product

$$\langle f, g \rangle \quad = \quad \int_{\Omega^n} f(x) g(x) d\mu^n(x). \tag{16}$$

Since $f, g \in F_n$ might not be in $L_2(\Omega^n, \mathcal{F}^n, \mu^n)$, the expression in (16) in general might not be finite. However, for $g \in L_\infty(\Omega^n, \mathcal{F}^n, \mu^n)$, we have

$$|\langle f, g\rangle| \quad \leq \quad \|f\|_1 \|g\|_\infty . \tag{17}$$

Define the *projection* operator $\pi : F_n \to F_{n-1}$ as follows. If $f : \Omega^n \to \mathbb{R}$ then $(\pi f) : \Omega^{n-1} \to \mathbb{R}$ is given by

$$(\pi f)(x_2, \ldots, x_n) \quad = \quad \int_\Omega f(x_1, x_2, \ldots, x_n) d\mu(x_1). \tag{18}$$

Note that by Fubini's theorem (Thm. 8.8(c) in Rudin (1987)), $\pi f \in L_1(\Omega^{n-1}, \mu^{n-1})$. Define the functional $\Psi_n : F_n \to \mathbb{R}$ recursively: $\Psi_0 = 0$ and

$$\Psi_n(f) \quad = \quad \Psi_{n-1}(\pi f) + \int_{\Omega^n} (f(x))_+ \, d\mu^n(x) \tag{19}$$

for $n \geq 1$. The latter is finite since

$$\Psi_n(f) \quad \leq \quad n \|f\|_{L_1(\mu)} , \tag{20}$$

as shown in the following Lemma.

Let $\Phi_n \subset F_n$ be the set of all measurable[4] $f : \Omega^n \to [0, \mathrm{diam}_\rho(\Omega^n)]$ with $\|f\|_{\mathrm{Lip}} \leq 1$ and define two norms on F_n:

$$\|f\|_\Phi \quad = \quad \sup_{g \in \Phi_n} |\langle f, g\rangle| \tag{21}$$

and

$$\|f\|_\Psi \quad = \quad \max_{s \in \{-1,1\}} \Psi_n(sf). \tag{22}$$

We refer to the norms in (21) and (22) as Φ-norm and Ψ-norm, respectively, and summarize some of their properties:

Lemma 3.1. *Under mild measure-theoretic regularity conditions on $(\Omega^n, \mathcal{F}, \mu^n)$, which cover the case of Ω countable and $\Omega = \mathbb{R}$ with Lebesgue measure (the metric being $\rho = d_{\mathrm{Ham}}$ in either case), the functionals defined in (21) and (22) satisfy*

(a) Φ-norm and Ψ-norm are valid vector-space norms on F_n

(b) for all $f \in F_n$,

$$\tfrac{1}{2} \|f\|_{L_1} \;\leq\; \|f\|_\Phi \;\leq\; n \|f\|_{L_1}$$

(c) for all $f \in F_n$,

$$\tfrac{1}{2} \|f\|_{L_1} \;\leq\; \|f\|_\Psi \;\leq\; n \|f\|_{L_1}$$

[4] Note that $\|f\|_{\mathrm{Lip}} \leq 1$ does not guarantee that f is $\mathcal{F}$-measurable, so the requirement that $\Phi_n \subset F_n$ is essential.

Proof. The claim in (a) is proved in (Kontorovich, 2006c, Theorems A.1, A.2, A.3); (b) and (c) are proved ibid. in (52) and Theorem A.1(b), respectively. $\square$

Definition 3.2. *A metric space* (Ω^n, ρ) *is said to be* Ψ-dominated *with respect to a positive Borel measure* μ *on* Ω *if the inequality*

$$\sup_{g \in \Phi_n} \langle f, g \rangle \quad \leq \quad \Psi_n(f) \tag{23}$$

holds for all $f \in F_n$.

Remark 3.3. We have defined Φ_n, $\Psi_n(\cdot)$, and $\langle \cdot, \cdot \rangle$ in an abstract measure space. To indicate explicitly what space they are being defined over, we will use the notation Φ_{Ω^n,μ^n}, $\Psi_{\Omega^n,\mu^n}(\cdot)$, and $\langle \cdot, \cdot \rangle_{\Omega^n,\mu^n}$.

When Ω is a finite set with counting measure ν, and $\rho = d_{\mathrm{Ham}}$, we have

$$\langle f, g \rangle_{\Omega^n,\nu^n} \quad \leq \quad \Psi_{\Omega^n,\nu^n}(f) \tag{24}$$

for all $f : \Omega^n \to \mathbb{R}$ and $g \in \Phi_{\Omega^n,\nu^n}$; in other words, $(\Omega^n, d_{\mathrm{Ham}})$ is Ψ-dominated with respect to ν. This was first proved in Kontorovich and Ramanan (2008); see Kontorovich (2006a) for a much simpler proof. Kontorovich (2007) extends this to countable Ω. The goal of the remainder of this section is to establish the analogue of (24) for $\Omega = \mathbb{R}$ with the Lebesgue measure μ.

Theorem 3.4. *Let* μ^n *be the Lebesgue measure on* $\mathbb{R}^n$ *and take* $\rho = d_{\mathrm{Ham}}$. *Then* $(\mathbb{R}^n, d_{\mathrm{Ham}})$ *is* Ψ*-dominated. That is, for all* $f \in L_1(\mathbb{R}^n, \mu^n)$ *and* $g \in \Phi_{\mathbb{R}^n,\mu^n}$, *we have*

$$\langle f, g \rangle_{\mathbb{R}^n,\mu^n} \quad \leq \quad \Psi_{\mathbb{R}^n,\mu^n}(f). \tag{25}$$

Remark 3.5. Observe that (25) is equivalent to

$$\|f\|_\Phi \quad \leq \quad \|f\|_\Psi \,, \qquad f \in L_1(\mathbb{R}^n, \mu^n). \tag{26}$$

The proof will closely follow the argument in Kontorovich (2006c, Theorem 8.1).

Proof. Let C_c denote the space of continuous functions $f : \mathbb{R}^n \to \mathbb{R}$ with compact support; it follows from (Rudin, 1987, Theorem 3.14) that C_c is dense in $L_1(\mathbb{R}^n, \mu^n)$, in the topology induced by $\|\cdot\|_{L_1}$. This implies that for any $f \in L_1(\mathbb{R}^n, \mu^n)$ and $\varepsilon > 0$, there is a $g \in C_c$ such that $\|f - g\|_{L_1} < \varepsilon/n$ and therefore (via Lemma 3.1(b) and (c)),

$$\|f - g\|_\Phi < \varepsilon \qquad \text{and} \qquad \|f - g\|_\Psi < \varepsilon$$

so it suffices to prove (26) for $f \in C_c$.

For $m \in \mathbb{N}$, define $Q_m \subset \mathbb{Q}$ to be the rational numbers with denominator m:

$$Q_m \quad = \quad \{p/r \in \mathbb{Q} : r = m\}\,.$$

Define the map $\gamma_m : \mathbb{R} \to Q_m$ by

$$\gamma_m(x) \quad = \quad \max\{q \in Q_m : q \leq x\}$$

and extend it to $\gamma_m : \mathbb{R}^n \to Q_m^n$ by defining $[\gamma_m(x)]_i = \gamma_m(x_i)$. The set $Q_m^n \subset \mathbb{R}^n$ will be referred to as the *m-grid points*.

We say that $g \in L_1(\mathbb{R}^n, \mu^n)$ is a *grid-constant function* if there is an $m > 1$ such that $g(x) = g(y)$ whenever $\gamma_m(x) = \gamma_m(y)$; thus a grid-constant function is constant on the grid cells induced by Q_m. Let G_c be the space of the grid-constant functions with compact support; note that $G_c \subset L_1(\mathbb{R}^n, \mu^n)$. It is easy to see that G_c is dense in C_c. Indeed, pick any $f \in C_c$ and let $M \in \mathbb{N}$ be such that $\mathrm{supp}(f) \subset [-M, M]^n$. Now a continuous function is uniformly continuous on a compact set, and so for any $\varepsilon > 0$, there is a $\delta > 0$ such that $\omega_f(\delta) < \varepsilon/(2M)^n$, where ω_f is the ℓ_∞ modulus of continuity of f. Take $m = \lceil 1/\delta \rceil$ and let $g \in G_c$ be such that $\mathrm{supp}(g) \subset [-M, M]^n$ and g agrees with f on the m-grid points. Then we have

$$\|f - g\|_{L_1} \leq (2M)^n \|f - g\|_{L_\infty} < \varepsilon.$$

Thus we need only prove (25) for $f \in G_c$, $g \in G_c \cap \Phi_{\mathbb{R}^n, \mu^n}$.

Let $f \in G_c$ and $g \in G_c \cap \Phi_{\mathbb{R}^n, \mu^n}$ be fixed, and let $m > 1$ be such that f and g are m-grid-constant functions. Let $\bar{\kappa}, \bar{\varphi} : Q_m^n \to \mathbb{R}$ be such that $\bar{\kappa}(\gamma_m(x)) = f(x)$ and $\bar{\varphi}(\gamma_m(x)) = g(x)$ for all $x \in \mathbb{R}^n$. As above, choose $M \in \mathbb{N}$ so that $\mathrm{supp}(f) \cup \mathrm{supp}(g) \subset [-M, M]^n$. Then, denoting the counting measure on Q_m^n by ν^n, we have

$$\langle f, g \rangle_{\mathbb{R}^n, \mu^n} = \left(\frac{2M}{m}\right)^n \langle \bar{\kappa}, \bar{\varphi} \rangle_{Q_m^n, \nu^n}$$

and

$$\Psi_{\mathbb{R}^n, \mu^n}(f) = \left(\frac{2M}{m}\right)^n \Psi_{Q_m^n, \nu^n}(\bar{\kappa}).$$

Now Q_m is finite and by construction, $\bar{\varphi} \in \Phi_{Q_m^n, \nu^n}$, so (24) applies. This shows $\langle f, g \rangle_{\mathbb{R}^n, \mu^n} \leq \Psi_{\mathbb{R}^n, \mu^n}(f)$ and completes the proof. □

3.2 Bounding the martingale difference

The machinery of η-mixing and Ψ-dominance allows us to bound the martingale difference for a Lipschitz function of arbitrarily dependent random variables.

Theorem 3.6. *Let $(\Omega, \mathcal{F}, \mu)$ be a positive Borel measure space and suppose that (Ω^n, ρ) is Ψ-dominated with respect to μ for some metric ρ on Ω^n. Let $(\Omega^n, \mathcal{F}^n, \mathbf{P})$ be a probability space with $\mathbf{P} \ll \mu^n$. Then for $1 \leq i \leq n$,*

$$\|V_i(f; \cdot)\|_{L_\infty(\mathbf{P})} \leq \|f\|_{\mathrm{Lip}} \|\Delta_n(\mathbf{P})\|_\infty, \tag{27}$$

where V_i is the martingale difference defined in (7) and Δ_n is the η-mixing matrix defined in (4).

Remark 3.7. This result can be proved, almost verbatim, by the argument given in Kontorovich (2006c, Theorem 7.1), so we only give a sketch of the proof here.

Proof. Since $\|V_i(f;\cdot)\|_{L_\infty}$ and $\|f\|_{\mathrm{Lip}}$ are both homogeneous functionals of f (in the sense of $T(af) = |a|T(f)$ for $a \in \mathbb{R}$), there is no loss of generality in taking $\|f\|_{\mathrm{Lip}} = 1$. Additionally, since $V_i(f;y)$ is translation-invariant (in the sense that $V_i(f;y) = V_i(f+a;y)$ for all $a \in \mathbb{R}$), there is no loss of generality in restricting the range of f to $[0,n]$. In other words, it suffices to consider $f \in \Phi_{\Omega^n,\mu^n}$.

From (12) we have that it suffices to bound $\|\hat{V}_i(f;\cdot)\|_{L_\infty}$, defined in (11). By (13), we have the equivalent form

$$\hat{V}_i(f;y_{1:i-1},w_i,w_i') \;=\; \int_{\Omega^n} f(x)\hat{g}(x)d\mu^n(x) \;=\; \langle f, g_i\rangle_{\Omega^n,\mu^n}\,, \tag{28}$$

where g_i has a simple explicit construction (14), depending on $y_{1:i-1}, w_i, w_i'$.

It is shown in the course of proving (Kontorovich, 2006c, Theorem 7.1) that

$$\langle f, g_i\rangle \;=\; \langle T_y f, T_y g_i\rangle\,,$$

where the operator $T_y : L_1(\mathbb{R}^n,\mu^n) \to L_1(\Omega^{n-i+1},\mu^{n-i+1})$ is defined by

$$(T_y h)(x) \;=\; h(yx), \qquad \text{for all } x \in \Omega^{n-i+1}.$$

Appealing to Theorem 3.4, we get

$$\langle T_y f, T_y g_i\rangle \;\le\; \Psi_n(T_y g_i). \tag{29}$$

Furthermore, as shown ibid., the form of g_i implies that

$$\Psi_n(T_y g_i) \;\le\; 1 + \sum_{j=i+1}^{n} \bar{\eta}_{ij}, \tag{30}$$

establishing (27).

□

4 The Strong Law of Large Numbers

We are now in a position to state our main result.

Theorem 4.1. *Let $(\Omega,\mathcal{F},\mu)$ be a positive Borel measure space and suppose that $(\Omega^n, d_{\mathrm{Ham}})$ is Ψ-dominated with respect to μ. Define the random process $X_{1:\infty}$ on the measure space $(\Omega^{\mathbb{N}},\mathcal{F}^{\mathbb{N}},\mathbf{P})$, and assume that for all $n \ge 1$ we have $\mathbf{P}_n \ll \mu^n$, where $\mathbf{P}_n$ is the marginal distribution on $X_{1:n}$ and μ^n is the corresponding product measure on $(\Omega^n,\mathcal{F}^n)$. Suppose further that the empirical measure defined by*

$$\hat{P}_n(A) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}, \quad A \in \mathcal{F}, \tag{31}$$

has uniformly converging expectation:

$$\lim_{n\to\infty} \left\| \mathbf{E}\hat{P}_n(\cdot) - \nu(\cdot) \right\|_{\mathrm{TV}} \quad \to \quad 0$$

and define $n_0 = n_0(\varepsilon)$ to be such that $\left\| \mathbf{E}\hat{P}_n(\cdot) - \nu(\cdot) \right\|_{\mathrm{TV}} < \varepsilon$ for all $n > n_0(\varepsilon)$.

Then $\hat{P}_n(A)$ converges to $\nu(A)$ almost surely, exponentially fast:

$$\mathbf{P}_n \left\{ \left| \hat{P}_n(A) - \nu(A) \right| > t + \varepsilon \right\} \quad \leq \quad 2\exp(-nt^2/2 \left\| \Delta_n \right\|_\infty^2) \tag{32}$$

for all $n > n_0(\varepsilon)$, where Δ_n is the η-mixing matrix defined in equation (4).

Particular cases of interest include Ω countable with μ taken to be counting measure and $\Omega = \mathbb{R}$ with μ taken to be Lebesgue measure.

Proof. This result follows directly from Theorem 3.6, by observing that the function $\varphi_A : X_{1:n} \to \mathbb{R}$ defined by $\varphi_A(X_{1:n}) = \hat{P}_n(A)$ has Lipschitz constant $1/n$. □

Corollary 4.2. *Under the conditions of Theorem 4.1, $\hat{P}_n$ converges to ν in distribution, almost surely.*

Proof. This is an immediate consequence of the (first) Borel-Cantelli lemma. □

5 Concentration of Marginals of Markov chains

For clarity of presentation, we take all the state spaces to be finite, until indicated otherwise in Section 5.4. Everything extends easily to the continuous case, as shown in the Appendix.

5.1 Bounding η_{ij} via contraction coefficients

Consider a Markov process

$$(W_t)_{t=1,2,\dots}$$

taking values $W_t = (X_t, Y_t) \in \Omega_{\mathrm{o}} \times \Omega_{\mathrm{h}}$, defined on the probability space $((\Omega_{\mathrm{o}} \times \Omega_{\mathrm{h}})^{\mathbb{N}}, (\mathcal{F}_{\mathrm{o}} \otimes \mathcal{F}_{\mathrm{h}})^{\mathbb{N}}, \mathbf{P})$; the subscripts 'o' and 'h' are used to suggest "observed" and "hidden" states. Suppose that we are interested primarily in the marginal behaviour of (X_t). (This might be the case, for instance, if we can observe (X_t) but not (Y_t) as when analyzing hidden Markov models. It is also of interest in the context of analysis of adaptive Markov chain Monte Carlo schemes as described later in this section.) Let us write the transition kernel of (W_t) as

$$K_t(w, A) = \mathbf{P}(W_{t+1} \in A \,|\, W_t = w), \quad w \in (\Omega_{\mathrm{o}} \times \Omega_{\mathrm{h}}), A \in (\mathcal{F}_{\mathrm{o}} \otimes \mathcal{F}_{\mathrm{h}}),$$

and assume that the initial distribution of the process is

$$p_0(A) = \mathbf{P}(W_1 \in A).$$

When $(X_t)_{t\in\mathbb{N}}$ can be embedded into a higher-dimensional Markov process as stated above, we will call (X_t) a *Markov marginal chain* (MMC). Note that Markov marginal chains properly contain the ordinary Markov chains, and are easily seen to be equivalent in expressive power to hidden Markov chains.

In this section we apply the results established in the previous sections to obtain relatively simple conditions for convergence of empirical measures associated with the MMC.

Let us define the i^{th} *contraction coefficient* θ_i of the MMC defined above by

$$\theta_i \quad = \quad \sup_{x^{\mathrm{o}},x^{\mathrm{o}\prime}\in\Omega_{\mathrm{o}}} \sup_{x^{\mathrm{h}},x^{\mathrm{h}\prime}\in\Omega_{\mathrm{h}}} \left\| K_i\left(\begin{array}{c} x^{\mathrm{o}} \\ x^{\mathrm{h}} \end{array}, \cdot\right) - K_i\left(\begin{array}{c} x^{\mathrm{o}\prime} \\ x^{\mathrm{h}\prime} \end{array}, \cdot\right) \right\|_{\mathrm{TV}}. \tag{33}$$

We obtain a bound on the η-mixing coefficients of the MMC in terms of its contraction coefficients:

Theorem 5.1. *The MMC (X_t) on $\Omega_{\mathrm{o}}^{\mathbb{N}}$, as defined above, satisfies*

$$\bar{\eta}_{ij} \quad = \quad \theta_i\theta_{i+1}\ldots\theta_{j-1}. \tag{34}$$

Proof. The proof is greatly simplified by an observation of Marton (2007) – namely, that a Marginal Markov chain is a special case of a Hidden Markov chain (see Rabiner (1989)). This is easily seen by considering the function

$$\pi : \left(\begin{array}{c} x^{\mathrm{o}} \\ x^{\mathrm{h}} \end{array}\right) \mapsto x^{\mathrm{o}},$$

which projects an (observed,hidden) pair onto its observed component, mapping a Markov chain to its hidden Markov marginal. It has already been shown (see Kontorovich (2006b) or Kontorovich and Ramanan (2008)) that the η-mixing coefficients of a hidden Markov chain are controlled by the contraction coefficients of the underlying Markov chain, in the manner of (34). □

5.2 Tensorization lemma

Lemma 5.2. *Suppose we have a MMC on $(\Omega_{\mathrm{o}}\times\Omega_{\mathrm{h}})^n$ defined by transition kernels $\{K_i(\cdot\,|\,\cdot)\}_{1\le i\le n}$, which have the following special structure:*

$$K_i\left(\begin{array}{c|c} u^{\mathrm{o}} & v^{\mathrm{o}} \\ u^{\mathrm{h}} & v^{\mathrm{h}} \end{array}\right) \quad = \quad A_i(u^{\mathrm{o}}\,|\,v^{\mathrm{o}},v^{\mathrm{h}})B_i(u^{\mathrm{h}}\,|\,v^{\mathrm{o}},v^{\mathrm{h}}). \tag{35}$$

Then we have

$$\theta_i \quad \le \quad \alpha_i+\beta_i-\alpha_i\beta_i \tag{36}$$

where θ_i is defined in (33), and

$$\begin{aligned}
\alpha_i &= \max_{x^{\mathrm{o}},x^{\mathrm{o}\prime}\in\Omega_{\mathrm{o}}} \max_{x^{\mathrm{h}},x^{\mathrm{h}\prime}\in\Omega_{\mathrm{h}}} \left\| A_i(\cdot \,|\, x^{\mathrm{o}}, x^{\mathrm{h}}) - A_i(\cdot \,|\, x^{\mathrm{o}\prime}, x^{\mathrm{h}\prime}) \right\|_{\mathrm{TV}}, \\
\beta_i &= \max_{x^{\mathrm{o}},x^{\mathrm{o}\prime}\in\Omega_{\mathrm{o}}} \max_{x^{\mathrm{h}},x^{\mathrm{h}\prime}\in\Omega_{\mathrm{h}}} \left\| B_i(\cdot \,|\, x^{\mathrm{o}}, x^{\mathrm{h}}) - B_i(\cdot \,|\, x^{\mathrm{o}\prime}, x^{\mathrm{h}\prime}) \right\|_{\mathrm{TV}}.
\end{aligned}$$

Proof. The claim follows immediately from the well-known total variation tensorization lemma (see for instance Kontorovich (2007)), which states that if μ, μ' are probability measures on $\mathcal{X}$ and ν, ν' are probability measures on $\mathcal{Y}$, then

$$\left\|\mu \otimes \nu - \mu' \otimes \nu'\right\|_{\mathrm{TV}} \quad \leq \quad \left\|\mu - \mu'\right\|_{\mathrm{TV}} + \left\|\nu - \nu'\right\|_{\mathrm{TV}} - \left\|\mu - \mu'\right\|_{\mathrm{TV}} \left\|\nu - \nu'\right\|_{\mathrm{TV}}$$

where $\mu \otimes \nu$ is a product measure on $\mathcal{X} \times \mathcal{Y}$. ☐

5.3 Concentration of adaptive Markov chains

Throughout our calculations, n will be a fixed positive integer. Let Γ be an index set and $\{K_\gamma(\cdot \,|\, \cdot)\}_{\gamma\in\Gamma}$ be a collection of Markov transition kernels, $K_\gamma : \Omega \to \Omega$. For a given sequence $\gamma_{1:n-1} \in \Gamma^{n-1}$, we have a Markov measure on Ω^n:

$$P_{\gamma_{1:n-1}}(x) \quad = \quad p_0(x_1) \prod_{i=1}^{n-1} K_{\gamma_i}(x_{i+1} \,|\, x_i), \qquad x \in \Omega^n.$$

For $1 \leq i < n$, $x \in \Omega$, and $\gamma' \in \Gamma$, let $g_i(\cdot \,|\, x, \gamma')$ be a probability measure on Γ. Together, $\{K_\gamma\}$ and $\{g_i\}$ define a measure μ on Ω^n, which we call an *adaptive Markov* measure:

$$\mu(x) \quad = \quad \sum_{\gamma_{1:n}\in\Gamma^n} p_0(x_1, \gamma_1) \prod_{i=1}^{n-1} [g_i(\gamma_{i+1} \,|\, x_i, \gamma_i) K_{\gamma_i}(x_{i+1} \,|\, x_{i+1})], \tag{37}$$

where γ_1 is a dummy index to make $g_1(\cdot \,|\, \cdot)$ well-defined.

Define the following contraction coefficients

$$\kappa \quad = \quad \max_{w,w'\in\Omega,\gamma,\gamma'\in\Gamma} \left\|K_\gamma(\cdot \,|\, w) - K_{\gamma'}(\cdot \,|\, w')\right\|_{\mathrm{TV}}. \tag{38}$$

and

$$\lambda_i \quad = \quad \max_{w,w'\in\Omega,\gamma,\gamma'\in\Gamma} \left\|g_i(\cdot \,|\, w, \gamma) - g_i(\cdot \,|\, w', \gamma')\right\|_{\mathrm{TV}}. \tag{39}$$

Then

Theorem 5.3. *For an adaptive Markov measure μ on Ω^n we have*

$$\bar{\eta}_{ij} \quad \leq \quad \theta_i \theta_{i+1} \cdots \theta_{j-1}$$

where

$$\theta_i = \kappa + \lambda_i - \kappa\lambda_i.$$

Proof. First we observe that the adaptive Markov process defined in (37) is in fact an MMC, with $\Omega_{\mathrm{o}} = \Omega$ and $\Omega_{\mathrm{h}} = \Gamma$; thus Theorem 5.1 applies. Furthermore, the MMC kernel (denoted here by P_i instead of K_i to avoid confusion with K_γ) decomposes as in (35):

$$P_i \left(\begin{array}{c|c} x & x' \\ \gamma & \gamma' \end{array} \right) \quad = \quad g_i(\gamma \,|\, x', \gamma') K_{\gamma'}(x \,|\, x'). \tag{40}$$

The claim is proved by applying Lemma 5.2 to (40). □

5.4 Example: Application to Adaptive MCMC Analysis

As one application, we consider the analysis of a family of adaptive Markov chain Monte Carlo schemes. Such schemes have been considered in some detail by a number of authors, beginning with studies of a specific scheme (Haario et al., 2001; Andrieu and Robert, 2001) and later being generalized by Atchade and Rosenthal (2005), Andrieu and Moulines (2006), and Roberts and Rosenthal (2007). Many other approaches to adaptation have been developed; these include, for example Doucet et al. (2000) and Brockwell and Kadane (2005). We consider the general framework of Roberts and Rosenthal (2007), and demonstrate that by imposing a stronger form of their so-called "diminishing adaptation" condition, one is able to strengthen the weak law of large numbers they establish to a strong law of large numbers.

Consider a stochastic process $\{X_t \in \mathbb{R},\ t = 0, 1, \ldots\}$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathbf{P})$. Adopting similar notation to that of Roberts and Rosenthal (2007), define a family $\{K_\gamma(\cdot,\cdot),\ \gamma \in \mathcal{G} \subseteq \mathbb{R}\}$ of transition kernels such that for each $\gamma \in \mathcal{G}$, $K_\gamma(\cdot,\cdot)$ is irreducible, aperiodic, and ergodic with a limiting distribution π on $(\mathbb{R}, \mathcal{F})$. One would typically take each such kernel to be a Metropolis-Hastings kernel with certain parameter values determined by γ.

For fixed γ, a homogeneous Markov chain whose joint distributions are determined by an initial value and transition kernel K_γ would have marginal distributions converging in total variation norm to π. However, in adaptive MCMC problems, interest centers on the behaviour of a more complex process $\{X_t\}$. Rather than holding γ fixed, one allows the transition kernel to vary over time. To be precise, we specify an initial value $X_0 = x_0$, along with transition probabilities

$$P(X_{t+1} \in A \,|\, X_t = x) = K_{\Gamma_t}(x, A),$$

where each $\Gamma_t \in \mathcal{G}$ is some function of $\Gamma_0, \ldots, \Gamma_{t-1}, X_0, \ldots, X_t$. Thus the kernel used at time t is itself random, depending on the past history of the process. This means that $\{X_t\}$ is not Markovian.

To see how we can apply Theorem 4.1 in this context, we introduce some assumptions.

Assumption 5.4. *There exists a random limiting kernel indexed by* $\Gamma_\infty(\omega)$ *such that*

$$\Gamma_t(\omega) \to \Gamma_\infty(\omega) \quad \forall \omega \in \Omega. \tag{41}$$

Furthermore, convergence to the random limit is uniform in the sense that there exists a non-negative monotone non-increasing sequence $\{\kappa_t\} \to 0$ *such that*

$$|\Gamma_t(\omega) - \Gamma_\infty(\omega)| \le \kappa_t \quad \forall \omega \in \Omega. \tag{42}$$

Assumption 5.5. *For each* $\gamma \in \mathcal{G}$*, we have* $K_\gamma(x, \cdot) \Rightarrow K_\gamma(y, \cdot)$ *(i.e., converges in distribution) whenever* $x \to y$*.*

Assumption 5.6. *Each kernel* $K_\gamma(\cdot, \cdot), \gamma \in \mathcal{G}$ *is uniformly ergodic, with*

$$\lim_{t\to\infty} \left\| K_\gamma^t(x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} = 0, \tag{43}$$

and satisfies the minorization condition

$$K_\gamma(x, \cdot) \ge m_0 \xi(\cdot), \quad \forall x \in \mathbb{R}, \tag{44}$$

where $\xi(\cdot)$ *is a probability measure on* $(\mathbb{R}, \mathcal{F})$ *and* m_0 *is some positive constant.*

Assumptions 5.4 and 5.6 are not restrictive. The first can be satisfied by requiring that $|\Gamma_{t+1} - \Gamma_t| = o(t^{-\alpha})$ for some $\alpha > 1$. Intuitively, this is just a form of what Roberts and Rosenthal (2007) refer to as "diminishing adaptation". Assumption 5.6 ensures that all possible adaptive kernels mix at a minimal rate, and that any composition of kernels also mixes at that rate. The latter follows from the well-known bound on the contraction coefficient θ via the minorization constant m_0 in (44):

$$\theta \le 1 - m_0;$$

see, for example, Lemma 2.2.3 in Kontorovich (2007). One way to construct such a family of kernels is to choose a family of Metropolis-Hastings kernels in which the proposal distributions all share a common component which does not depend on the current state.

Assumption 5.5 is a natural Feller-type condition; in particular, it is satisfied by most Metropolis-Hastings chains, and is fairly easily checked in practice.

Using our main result along with these assumptions, we are in a position to state conditions under which Theorem 23 of Roberts and Rosenthal (2007) can be strengthened to establish strong instead of weak convergence.

Theorem 5.7. *Suppose that an adaptive Markov chain satisfies Assumptions 5.4, 5.6, and 5.5. Then we have*

$$\hat{P}_n(A) \to \pi(A), \text{ a.s.},$$

for each $A \in \mathcal{F}$*, where* $\hat{P}_n(\cdot)$ *is the empirical measure defined in (31).*

Proof. The minorization condition in Assumption 5.6 ensures that the simultaneous uniform ergodicity condition of Theorem 5 of Roberts and Rosenthal (2007) holds. Assumption 5.4 ensures that the diminishing adaption condition of the same theorem is also satisfied. Then we have

$$\left|\mathbf{E}\hat{P}_n(A) - \pi(A)\right| \leq \epsilon + \mathbf{P}\left\{\left|\hat{P}_n(A) - \pi(A)\right| > \epsilon\right\}, \tag{45}$$

and the latter probability is bounded by a quantity independent of A, which goes to zero, by Theorem 23 of Roberts and Rosenthal (2007) (the uniformity is not stated explicitly in the Theorem but is established in the proof that they give). This establishes the uniformly converging expectation condition of our Theorem 4.1. The minorization condition (44) ensures that the mixing coefficients $\bar{\eta}_{ij}$ decay as $(1 - m_0)^{j-i}$. To establish almost-sure convergence, we are going to argue along the lines of Theorem 4.1 (the latter is not applicable directly, since the adaptive Markov chain measure P might not have a density with respect to the Lebesgue measure). Let $n_0 = n_0(\varepsilon)$ be such that $\left|\mathbf{E}\hat{P}_n(A) - \pi(A)\right| < \varepsilon$ for all $n > n_0(\varepsilon)$. As in Theorem 4.1, the quantity we want to bound is

$$P\left(\left|\hat{P}_n(A) - \pi(A)\right| > t + \varepsilon\right) \tag{46}$$

for $\varepsilon, t > 0$ and $n > n_0(\varepsilon)$.

We now approximate the chain P by a finite-state chain P' induced by finite partitions, as shown in the Appendix. It follows from the arguments in the Appendix that for any $E \in \mathcal{F}^n$ we can find a partition of the state space and a finite-state chain P' on $(\Omega', \mathcal{F}')^n$ so that $P'(\tilde{E})$ is arbitrarily close to $P(E)$ for some $\tilde{E} \in (\mathcal{F}')^n$. Therefore, $\hat{P}_n(A)$ and $\pi(A)$ can be made arbitrarily close to their finite-state analogues $\hat{P}'_n(\tilde{A})$ and $\pi'(\tilde{A})$ and in particular,

$$P'\left(\left|\hat{P}'_n(\tilde{A}) - \pi'(\tilde{A})\right| > t + \varepsilon\right) \tag{47}$$

approximates the expression in (46). Furthermore, Lemma A.1 shows that for a sufficiently refined partition, the finite-state chain P' will have mixing coefficients arbitrarily close to those of P. To conclude the proof it suffices to apply Theorem 4.1 and Corollary 4.2. □

Acknowledgements

We would like to thank Ofer Zeitouni for helpful discussions.

A Extending finite-state inequalities to more general spaces

In bounding the mixing coefficients, measure-theoretic technicalities tend to play a peripheral role. Indeed, the $\Omega = \{0, 1\}$ case already captures most of the proof complexity. The mixing results we proved for finite Ω extend verbatim to $\Omega = \mathbb{N}$, and under mild continuity assumptions, to much more general measure spaces.

The inequality in Theorem 3.6 applies to a broad class of state spaces, including $\Omega = \mathbb{R}$. The latter result estabishes that the η-mixing coefficients of a random process control the concentration of Lipschitz path functionals about their means. This immediately implied the strong law of large numbers in Section 4.

In Theorem 5.3 we showed how to control the η_{ij} in terms if the Markov contraction coefficients; this was done for finite state spaces. In this Appendix, we extend these bounds to the continuous case.

A.1 Markov marginal chains

In the continuous case we shall consider the "observed" state space Ω_{o} and the "hidden" state space Ω_{h}, with the total space $\mathcal{W} = \Omega_{\mathrm{o}} \times \Omega_{\mathrm{h}}$; one may take $\Omega_{\mathrm{o}} = \Omega_{\mathrm{h}} = \mathbb{R}$. We take the usual Borel σ-algebra on $\mathcal{W}$, which is a product of the corresponding σ-algebras on Ω_{o} and Ω_{h}. An MMC over Ω_{o}^n is obtained by first defining the Markov process $W_t = (X_t, Y_t)$, with values in $\mathcal{W}$, induced by the kernels $\{K_i\}_{0 \le i < n}$. Thus if $A_i \subset \mathcal{W}$, $1 \le i \le n$ are measurable, then

$$\mathbf{P}\{W_1 \in A_1, \ldots, W_n \in A_n\} \;=\; \int_{A_1} \cdots \int_{A_n} \prod_{i=1}^{n} K_i(w_{i-1}, dw_i).$$

We will write K_0 for the initial distribution of the MMC, and expressions such as $K_0(x_0, dx_1)$ are to be interpreted as $K_0(dx_1)$. Note that we are not assuming anything about the density of P; the latter may well not exist with respect to the Lebesgue (or any other product) measure on $\mathcal{W}^n$.

As in the discrete case, the MMC is obtained by marginalizing out the "hidden" component Y:

$$\mathbf{P}\{X_1 \in B_1, \ldots, X_n \in B_n\} \;=\; \int_{\Omega_{\mathrm{h}}^n} \int_{B_1} \cdots \int_{B_n} \prod_{i=1}^{n} K_i((x_{i-1}, y_{i-1}), (dx_i, dy_i))$$

for measurable sets $B_i \subset \Omega_{\mathrm{o}}$.

The definition of the contraction coefficient readily generalizes to the continuous case:

$$\theta_i \;=\; \sup_{w, w' \in \mathcal{W}} \left\| K_i(w, \cdot) - K_i(w', \cdot) \right\|_{\mathrm{TV}}. \tag{48}$$

We will use the notation $\theta_i(P)$ if we wish to make explicit the dependence on the particular Markov measure.

Similarly, we define

$$\bar{\eta}_{ij} \;=\; \sup_{y \in \mathcal{W}^{i-1}, w, w' \in \mathcal{W}} \eta_{ij}(y, w, w'),$$

and write $\bar{\eta}_{ij}(P)$ to make explicit the particular measure.

All sets below are assumed to be Borel-measurable. Recall that a *partition* of a set E is a collection of disjoint sets whose union is E. Whenever $\mathcal{A}$ is a collection of subsets of E, $\mathcal{A}$ induces an equivalence

relation on E as follows:

$$x \equiv_{\mathcal{A}} y \iff \mathbb{1}_{\{x \in A\}} = \mathbb{1}_{\{y \in A\}} \text{ for all } x, y \in E, A \in \mathcal{A}.$$

Such an equivalence relation in turn induces a partition on E (whose members are the equivalence classes); we shall call such a partition the *refining partition* of $\mathcal{A}$. Notice that if $\mathcal{A}$ is finite then so is its refining partition.

Let P be a Markov chain on $\mathcal{W}^n$, as above. Any finite partition $\{V_k : 1 \leq k \leq m\}$ of $\mathcal{W}$ induces a Markov chain $\hat{P}$ with kernels $\left\{\hat{K}_i\right\}_{0 \leq i < n}$ over the finite state space $\hat{\mathcal{W}} = \{1, \ldots, m\}$, as follows:

$$\hat{K}_i(k', k) \quad = \quad P(W_{i+1} \in V_k \,|\, W_i \in V_{k'}).$$

If $\mathcal{W} = \Omega_{\mathrm{o}} \times \Omega_{\mathrm{h}}$, $\{S_k\}$ is a partition of Ω_{o}, and $\{T_\ell\}$ is a partition of Ω_{h}, then these two partitions induce a partition on $\mathcal{W}$ in the obvious way. We will write $\hat{\mathcal{W}} = \hat{\Omega_{\mathrm{o}}} \times \hat{\Omega_{\mathrm{h}}}$ to denote the state spaces obtained by identifying the partition blocks with states.

Lemma A.1. *Let P be a Markov chain on $\mathcal{W}^n = (\Omega_{\mathrm{o}} \times \Omega_{\mathrm{h}})^n$ and Q be the induced MMC on Ω_{o}^n. Then, assuming that the kernel generating P satisfies the Feller continuity condition in Assumption 5.5, we have that for any $\varepsilon > 0$ there are finite partitions of Ω_{o} and Ω_{h} such that the induced Markov chain $\hat{P}$ on $\hat{\mathcal{W}}^n = (\hat{\Omega_{\mathrm{o}}} \times \hat{\Omega_{\mathrm{h}}})^n$ and MMC $\hat{Q}$ on $\hat{\Omega_{\mathrm{o}}}^n$ satisfy*

$$\left|\theta_i(P) - \theta_i(\hat{P})\right| \quad < \quad \varepsilon \tag{49}$$

and

$$\left|\bar{\eta}_{ij}(Q) - \bar{\eta}_{ij}(\hat{Q})\right| \quad < \quad \varepsilon, \tag{50}$$

for $1 \leq i < j \leq n$.

Proof. Fix an $\varepsilon > 0$. Let us construct the requisite partition for (49). Fix an $1 \leq i < n$ and let $t = \theta_i(P)$. Then by definition of $\|\cdot\|_{\mathrm{TV}}$ there are w and w' in $\mathcal{W}$, and an $E \subset \mathcal{W}$ such that

$$t - \varepsilon/2 \;\leq\; K_i(w, E) - K_i(w', E) \;\leq\; t.$$

Furthermore, E may be approximated by a finite union of rectangles $A_k \times B_\ell$, with $A_k \subset \Omega_{\mathrm{o}}$, $k = 1, \ldots, m_A$, and $B_\ell \subset \Omega_{\mathrm{h}}$, $\ell = 1, \ldots, m_B$:

$$\tilde{E} \quad = \quad \bigcup_{k=1,\ldots,m_A} \bigcup_{\ell=1,\ldots,m_B} A_k \times B_\ell, \tag{51}$$

so that

$$t - \varepsilon \;\leq\; K_i(w, \tilde{E}) - K_i(w', \tilde{E}) \;\leq\; t.$$

By Feller continuity, we have

$$K_i(w, E) \quad = \quad \lim_{\alpha \to \infty} P(W_{i+1} \in E \,|\, W_i \in U_\alpha)$$

where $U_1 \supset U_2 \supset \ldots \ni w$ is a sequence of neighborhoods *shrinking to* w in the sense that $\cap_\alpha U_\alpha = w$. (We are assuming without loss of generality that $P(U_\alpha) > 0$ for all α.)

Therefore, taking $w = (x, y)$ and $w' = (x', y')$, we have that there are neighborhoods $C, C' \subset \Omega_\mathrm{o}$ of x and x' (respectively), as well as analogous neighborhoods $D, D' \subset \Omega_\mathrm{h}$ of y and y', such that

$$P(W_{i+1} \in E \,|\, W_i \in C \times D) - P(W_{i+1} \in E \,|\, W_i \in C' \times D')$$

is arbitrarily close to $K_i(w, \tilde{E}) - K_i(w', \tilde{E})$.

Let $\{S_a\}$ be the partition of Ω_o refining the sets $\{A_k\}$, C, and C'. Similarly, let $\{T_b\}$ be the partition of Ω_h refining the sets $\{B_\ell\}$, D, and D'. Identify the partition blocks with states and induce the finite-state Markov chain $\hat{P}$ on $\hat{\mathcal{W}} = \hat{\Omega_\mathrm{o}} \times \hat{\Omega_\mathrm{h}}$. By the approximation argument above, there exist C, C', D, D' such that $\hat{P}$ satisfies (49) for the given i. Repeating this process for each $i = 1, \ldots, n-1$ and picking a finite partition of Ω_o (respectively, Ω_h) that simultaneously refines the partitions for each i, we establish (49) for all i.

Now we turn to (50). The basic technique is the same as the one used to show (49). Fix $1 \le i < j \le n$ and let $h = \bar{\eta}_{ij}(Q)$. Then there are $x_{1:i-1} \in \Omega_\mathrm{o}^{i-1}$, $x_i, x_i' \in \Omega_\mathrm{o}$ and an $A \subset \Omega_\mathrm{o}^{n-j+1}$ such that

$$h - \varepsilon/2 \;\le\; Q(X_{j:n} \in A \,|\, X_{1:i} = x_{1:i}) - Q(X_{j:n} \in A \,|\, X_{1:i} = x_{1:i-1}x') \;\le\; h.$$

As done with E and $\tilde{E}$ above, $Q(X_{j:n} \in A \,|\, X_{1:i} = x_{1:i})$ may be approximated by $Q(X_{j:n} \in \tilde{A} \,|\, X_{1:i} = x_{1:i})$ where $\tilde{A}$ is now an $(n-j+1)$-fold product of expressions of the type $\cup_k A_k$, with $A_k \subset \Omega_\mathrm{o}$.

Again as above, we take small neighborhoods around $x_1, x_2, \ldots, x_i, x_i'$ to obtain a partition which satisfies (50) for the given i, j. Refining such partitions simultaneously for $1 \le i < j \le n$, we establish (50) for all these values.

Finally, the partitions obtained in the course of proving (49) and (50) can once again be refined to make the two inequalities hold simultaneously. □

Corollary A.2. *Let P be a Markov chain on $\mathcal{W}^n = (\Omega_\mathrm{o} \times \Omega_\mathrm{h})^n$ and Q be the induced MMC on Ω_o^n. Then*

$$\bar{\eta}_{ij}(Q) \;\; \le \;\; \prod_{k=i}^{j-1} \theta_k(P).$$

Proof. Immediate consequence of the corresponding claim for finite Ω_o and Ω_h. □

A.2 Tensorization

Lemma A.3. *Suppose the Markov chain P on $(\Omega_\mathrm{o} \times \Omega_\mathrm{h})^n$ is defined by transition kernels $\{K_i\}_{0 \le i \le n}$, which have the following special structure:*

$$K_i\left((x', y'), (x, y)\right) \;\; = \;\; A_i(x \,|\, (x', y')) B_i(y \,|\, (x', y')). \tag{52}$$

Then we have

$$\theta_i \leq \alpha_i + \beta_i - \alpha_i \beta_i \tag{53}$$

where

$$\begin{aligned}
\alpha_i &= \sup_{(x',y'),(x'',y'') \in \Omega_{\mathrm{o}} \times \Omega_{\mathrm{h}}} \left\| A_i(\cdot \,|\, (x',y')) - A_i(\cdot \,|\, (x'',y'')) \right\|_{\mathrm{TV}}, \\
\beta_i &= \sup_{(x',y'),(x'',y'') \in \Omega_{\mathrm{o}} \times \Omega_{\mathrm{h}}} \left\| B_i(\cdot \,|\, (x',y')) - B_i(\cdot \,|\, (x'',y'')) \right\|_{\mathrm{TV}}.
\end{aligned}$$

Proof. The same technique of approximating P by a finite-state Markov chain $\hat{P}$ as employed in Lemma A.1 may be used here; details are omitted. □

A.3 Adaptive Markov chains

Let Γ be an index set and $\{K_\gamma\}_{\gamma \in \Gamma}$ be a collection of Markov transition kernels, $K_\gamma : \Omega \to \Omega$. For a given sequence $\gamma_{1:n-1} \in \Gamma^{n-1}$, we have a Markov measure on Ω^n:

$$dP_{\gamma_{1:n-1}}(x) = \prod_{i=0}^{n-1} K_{\gamma_i}(x_i, dx_{i+1}), \qquad x \in \Omega^n.$$

For $1 \leq i < n$, $x \in \Omega$, and $\gamma' \in \Gamma$, let $g_i(\cdot \,|\, x, \gamma')$ be a probability measure on Γ. Together, $\{K_\gamma\}$ and $\{g_i\}$ define a measure Q on Ω^n, which we'll call an *adaptive Markov* measure:

$$dQ(x) = \int_{\Gamma^n} \prod_{i=1}^{n-1} [g_i(d\gamma_{i+1} \,|\, x_i, \gamma_i) K_{\gamma_i}(x_i, dx_{i+1})], \tag{54}$$

where γ_1 is a dummy index to make $g_1(\cdot \,|\, \cdot)$ well-defined.

Define the following contraction coefficients

$$\kappa = \sup_{w,w' \in \Omega, \gamma, \gamma' \in \Gamma} \left\| K_\gamma(w, \cdot) - K_{\gamma'}(w', \cdot) \right\|_{\mathrm{TV}}. \tag{55}$$

and

$$\lambda_i = \sup_{w,w' \in \Omega, \gamma, \gamma' \in \Gamma} \left\| g_i(\cdot \,|\, w, \gamma) - g_i(\cdot \,|\, w', \gamma') \right\|_{\mathrm{TV}}. \tag{56}$$

Then

Theorem A.4. *For an adaptive Markov measure Q on Ω^n we have*

$$\bar{\eta}_{ij} \quad \leq \quad \theta_i \theta_{i+1} \cdots \theta_{j-1}$$

where

$$\theta_i = \kappa + \lambda_i - \kappa\lambda_i.$$

Proof. First we observe that the adaptive Markov process defined in (54) is in fact a MMC, with $\Omega_{\mathrm{o}} = \Omega$ and $\Omega_{\mathrm{h}} = \Gamma$; thus Corollary A.2 applies. Furthermore, the MMC kernel (denoted here by P_i instead of K_i to avoid confusion with K_γ) decomposes as in (52):

$$P_i\left((x', \gamma'), (dx, d\gamma)\right) \quad = \quad g_i(d\gamma \,|\, x', \gamma') K_{\gamma'}(x', dx). \tag{57}$$

The claim is proved by applying Lemma A.3 to (57). □

References

C. Andrieu and E. Moulines. On the ergodicity properties of some adaptive mcmc algorithms. *Annals of Applied Probability*, 16(3):1462–1505, 2006.

C. Andrieu and C. Robert. Controlled mcmc for optimal sampling. Technical report, 2001. URL `citeseer.ist.psu.edu/andrieu01controlled.html`.

Y. F. Atchade and J. S. Rosenthal. On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11:815–828, 2005.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Math. Journal*, 19:357–367, 1967.

Jakob Bernoulli. *Ars Conjectandi: Usum & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis*, volume 4. 1713.

Richard C. Bradley. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probab. Surveys*, 2:107–144, 2005.

Leo Breiman. The strong law of large numbers for a class of Markov chains. *Ann. Math. Statist.*, 31:801–803, 1960.

Anthony E. Brockwell and Joseph B. Kadane. Identification of regeneration times in mcmc simulation, with application to adaptive schemes. *Journal of Computational & Graphical Statistics*, 14 (2):436–458, June 2005. ISSN 1061-8600. URL `http://dx.doi.org/10.1198/106186005X47453`.

Sourav Chatterjee. *Concentration inequalities with exchangeable pairs.* PhD thesis, Stanford University, 2005.

Jean-René Chazottes, Pierre Collet, Christof Külske, and Frank Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137(1-2):201–225, 2007.

Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000. ISSN 0960-3174.

Paul Doukhan and Sana Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84(2):313–342, 1999.

H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

David L. Hanson, Julius R. Blum, and Lambert H. Koopmans. On the strong law of large numbers for a class of stochastic processes. *Probability Theory and Related Fields*, 2(1):1–11, 1963.

Leonid Kontorovich. A Linear Programming Inequality with Applications to Concentration of Measure. 2006a.

Leonid Kontorovich. Measure Concentration of Hidden Markov Processes. 2006b.

Leonid Kontorovich. Metric and Mixing Sufficient Conditions for Concentration of Measure. 2006c.

Leonid Kontorovich. *Measure Concentration of Strongly Mixing Processes with Applications*. PhD thesis, Carnegie Mellon University, 2007.

Leonid Kontorovich. Constructing processes with prescribed mixing coefficients, to appear in *Statistics & Probability Letters*. 2008.

Leonid Kontorovich and Kavita Ramanan. Concentration Inequalities for Dependent Random Variables via the Martingale Method, to appear in *Ann. Probab.* 2008.

Michel Ledoux. *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs Vol. 89*. American Mathematical Society, 2001.

Gábor Lugosi. Concentration-of-measure inequalities, `http://www.econ.upf.es/~lugosi/anu.ps`. 2003.

Katalin Marton. Measure concentration for a class of random processes. *Probability Theory and Related Fields*, 110(3):427–439, 1998.

Katalin Marton, 2007. private communication.

Brett Ninness. Strong Laws of Large Numbers Under Weak Assumptions with Applications. *IEEE Transactions on Automatic Control*, 45(11):2117–2122, 2000.

Andrew Nobel. A counterexample concerning uniform ergodic theorems for a class of functions. *Statistics and Probability Letters*, 24:165–168, 1995.

Andrew Nobel and Amir Dembo. A note on uniform laws of averages for dependent processes. *Statistics and Probability Letters*, 17:169–172, 1993.

David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

Gareth O. Roberts and Jeffrey S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.*, 44(2):458–475, 2007. ISSN 0021-9002.

Walter Rudin. *Real and Complex Analysis.* McGraw-Hill, 1987.

Paul-Marie Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

Gideon Schechtman. *Concentration, results and applications. Handbook of the Geometry of Banach Spaces, Volume 2.* North-Holland, 2003.